ELSEVIER

Potential for Gene–Gene Confounding Bias in Case–Parental Control Studies

WEN-CHUNG LEE, MD, PHD, AND YEN-YI HO, MS

PURPOSE: To show the potential for gene–gene confounding bias in case–parental control studies.
METHODS: The authors quantify the magnitude of gene–gene confounding bias using simple mathematical equations. They also demonstrate the potential problems of such a bias with hypothetical (but realistic) examples.
RESULTS: The degree of bias resulting from gene–gene confounding was found to be quite substantial under certain conditions (two genes are very closely linked and/or the study was performed in a recently admixed population).
CONCLUSION: In this post–genomic era more and more encounters of the gene–gene confounding will be expected, if the one-gene-at-a-time approach continues to be adopted.
Ann Epidemiol 2003;13:261–266.

KEY WORDS: Case–Parental Control Study, Confounding, Epidemiologic Methods, Genetic Epidemiology, Genotype Relative Risk, Transmission/Disequilibrium Test.

INTRODUCTION

Genetic factors contribute to virtually every human disease, conferring susceptibility or resistance. Unlike the simple Mendelian diseases, many genes may be involved in the pathogenesis of the more common "complex" human diseases. For example, the susceptibility genes for breast cancer include *BRCA1*, *BRCA2*, *p53*, *HRAS1*, *HER-2/neu*, *COMT*, *CYP17*, *apoE*, *CYP1A1*, *GSTT1*, *GSTM1*, *NAT2*, *XRCC1*, *ATM*, *SNCG*, etc. (1–6).

To quantify the relation between susceptibility genes and disease risk, a novel study design, the "case–parental control study", has come to much attention (7, 8). Under this design, one needs only collect the genotype data of the cases and their parents, whereas a control group in the usual sense is not needed. Because the comparison is within family (conditional on parental genotype) (9, 10), the case–parental control study is not affected by population structure (11–13), which can otherwise confound the "genotype relative risk" (GRR) estimation in a conventional case–control study. Furthermore, there is no need to worry about possible confounding from environmental factors either, since the case–parental control design is a self-matched study with the "controls" having exactly the same environmental exposures as the case (14). (The "controls" in the case–parental control study are the parental untransmitted alleles.)

However, the estimation of the GRR for a particular susceptibility gene can be confounded by other genes even in a case–parental control study. Such "gene–gene confounding" can arise when several susceptibility genes have been genotyped in a study, but the analysis is performed on a one-gene-at-a-time basis. It can also arise when some important susceptibility gene(s) other than the one under study have not yet been identified and thus could not be genotyped and adjusted in the study. (It is important not to confuse gene–gene confounding with "gene–gene interaction" (15), a term that describes the effect modification of a gene by the presence of another gene. The gene–gene confounding operates at a more fundamental level. It is a bias to be reckoned with even if two genes act in a way predictable by the simplest multiplicative model.)

METHODS AND RESULTS

In this study, we will quantify using some simple mathematical equations, the magnitude of gene–gene confounding bias in a case–parental control study. We will also demonstrate the potential problems of such a bias with hypothetical (but realistic) examples.

Gene–Gene Confounding

Assume that two genetic loci, A and B, confer susceptibility to the risk of a disease.

From the Graduate Institute of Epidemiology, College of Public Health, National Taiwan University, Republic of China (W.-C.L., Y.-Y.H.).
Address correspondences to: Dr. Wen-Chung Lee, Graduate Institute of Epidemiology, National Taiwan University, No. 1, Jen-Ai Rd, 1st Sec, Taipei, Taiwan, Republic of China. Fax: +13-886-2-23511955. E-mail: wenchung@ha.mc.ntu.edu.tw

360 Park Avenue South, New York, NY 10010
1047-2797/03/$–see front matter
PII S1047-2797(02)00410-6

AEP Vol. 13, No. 4
April 2003: 261–266

There are two alleles at either locus, *A* (high risk) and *a* (low risk) alleles for A locus, and *B* (high risk) and *b* (low risk) alleles for B locus. Let $R_1{}^A$ denote the GRR of *Aa* genotype versus *aa* genotype, and let R_2^A, the GRR of *AA* versus *Aa*. The corresponding GRRs at B locus are denoted by R_1^B and R_2^B, respectively. Let R_0 represent the disease risk of subjects with genotype *aabb*. We assume there is no gene–gene interaction between A and B (in a multiplicative model). Therefore the disease risk for subjects with genotype *Aabb* is $R_0 \times R_1{}^A$, and the disease risk for *AaBB* is $R_0 \times R_1^A \times R_1^B \times R_2^B$, etc.

We assume that the A gene has been genotyped and is the gene under concern, and the B gene is the "confounder gene" (for whatever reason described above). We set out to examine the bias in the estimation of GRRs for the A gene, assuming that the B gene was not accounted for in a case–parental control study. The index of "relative bias" (RB) is used to quantify the magnitude of bias in proportional term [RB = (estimated value − true value)/true value]. It has a simple relation to the "confounding risk ratio" (CRR) used in a previous study (13), that is, RB = CRR − 1.

Unstructured Population

Assuming that the case–parental control study was performed in a homogeneous unstructured population (a random mating population in Hardy–Weinberg equilibrium) with population size of *N* subjects, the appendix shows that the relative bias for R_1^A (if not accounting for B) is

$$\mathrm{RB}_1 = (1-2\theta)\cdot\delta\cdot[Q_{aB}R_1^B(R_2^B-1) + Q_{ab}(R_1^B-1)]/F_1, \tag{1}$$

and the relative bias for R_2^A is

$$\mathrm{RB}_2 = (1-2\theta)\cdot\delta\cdot[Q_{AB}R_1^B(R_2^B-1) + Q_{Ab}(R_1^B-1)]/F_2, \tag{2}$$

where θ is the recombination fraction between A and B, δ is the linkage disequilibrium parameter in parental population, Q_{AB}, Q_{Ab}, Q_{aB}, and Q_{ab} are the haplotype frequencies in offspring for *AB*, *Ab*, *aB*, and *ab*, respectively, and $F_1 > 0$, $F_2 > 0$ are some complicated functions detailed in appendix. Note that

$$\delta = P_{AB}\cdot P_{ab} - P_{Ab}\cdot P_{aB} = (Q_{AB}\cdot Q_{ab} - Q_{Ab}\cdot Q_{aB})/(1-\theta), \tag{3}$$

where P_{AB}, P_{Ab}, P_{aB}, and P_{ab} are the haplotype frequencies in parental population (16). When the B gene displays a multiplicative gene-dose effect, that is $R_1^B = R_2^B = R^B$, the above equations for the relative biases in the estimation of the effects of A gene reduce to

$$\mathrm{RB}_1 = \mathrm{RB}_2 = (1-2\theta)\cdot\delta\cdot(R^B-1)/F, \tag{4}$$

where $F > 0$ is detailed in appendix.

From these equations, it is clear that we will obtain an unbiased estimation of GRRs for the A gene if either one of the following three conditions is fulfilled: (i) the B gene and the A gene are unlinked ($\theta = 0.5$); (ii) the B gene and the A gene are in the same chromosome but are separated widely apart, say >∼10cM, such that the two genes are in linkage equilibrium ($\delta = 0$); (iii) the B gene is not a susceptibility gene for the disease under study ($R_1^B = R_2^B = 1$). In other words, when a case–parental control study was performed in an unstructured population, nearby susceptibility gene(s) at the same chromosome could exert a confounding bias on GRR estimation of the study gene. The direction and magnitude of the bias is to our expectation, that is, smaller θ, larger δ, and/or larger GRRs (for B gene) will lead to larger bias and that the bias will be in a positive direction (overestimation) if A and B genes are positively correlated ($\delta > 0$) and in a negative direction (underestimation) if otherwise. It is of interest to see that the RBs for A gene do not depend on the GRRs of A gene itself.

Stratified Population

The study population is now assumed to be composed of two subpopulations (the first subpopulation constitutes *m* ($0 < m < 1$) proportion, and the second, $1 - m$). Random mating occurs within the subpopulations but the two subpopulations do not intermix. It can be shown that the relative biases in this stratified population at large are the weighted averages of the relative biases in the two subpopulations, that is (the superscripts, I and II, indicate the two subpopulations):

$$\begin{aligned}\mathrm{RB}_1^S &= \frac{mF_1^I\cdot\mathrm{RB}_1^I + (1-m)F_1^{II}\cdot\mathrm{RB}_1^{II}}{mF_1^I + (1-m)F_1^{II}} \\ &= (1-2\theta)\cdot \\ &\{m\delta^I\cdot[Q_{aB}^I R_1^B(R_2^B-1) + Q_{ab}^I(R_1^B-1)] + \\ &(1-m)\delta^{II}\cdot[Q_{aB}^{II}R_1^B(R_2^B-1) + Q_{ab}^{II}(R_1^B-1)]\}/ \\ &[mF_1^I + (1-m)F_1^{II}]\end{aligned} \tag{5}$$

and

$$\mathrm{RB}_2^{\mathrm{S}} = \frac{mF_2^{\mathrm{I}} \cdot \mathrm{RB}_2^{\mathrm{I}} + (1-m)F_2^{\mathrm{II}} \cdot \mathrm{RB}_2^{\mathrm{II}}}{mF_2^{\mathrm{I}} + (1-m)F_2^{\mathrm{II}}} = (1-2\theta) \cdot \{m\delta^{\mathrm{I}} \cdot [Q_{AB}^{\mathrm{I}} R_1^{\mathrm{B}}(R_2^{\mathrm{B}}-1) + Q_{Ab}^{\mathrm{I}}(R_1^{\mathrm{B}}-1)] + (1-m)\delta^{\mathrm{II}} \cdot [Q_{AB}^{\mathrm{II}} R_1^{\mathrm{B}}(R_2^{\mathrm{B}}-1) + Q_{Ab}^{\mathrm{II}}(R_1^{\mathrm{B}}-1)]\}/[mF_2^{\mathrm{I}} + (1-m)F_2^{\mathrm{II}}]. \quad (6)$$

It is then clear that in the stratified population, we will obtain unbiased estimations if either one of the following three conditions is fulfilled: (i) the B gene and the A gene are unlinked ($\theta = 0.5$); (ii) the B gene and the A gene are in the same chromosome but are separated widely apart, say >∼10cM, such that the two genes are in linkage equilibrium in both subpopulations ($\delta^{\mathrm{I}} = \delta^{\mathrm{II}} = 0$); (iii) the B gene is not a susceptibility gene for the disease under study ($R_1^{\mathrm{B}} = R_2^{\mathrm{B}} = 1$). These conditions are essentially the same as those in the case of unstructured population as described before, that is, only nearby susceptibility gene(s) can exert a confounding effect on the study gene. Note however that in general $\mathrm{RB}_1^{\mathrm{S}} \neq \mathrm{RB}_2^{\mathrm{S}}$ even with a multiplicative gene-dose effect for the B gene.

Admixed Population

Suppose that an admixed population obtains a fraction m of its genes from ancestral population I and a fraction $1 - m$ from ancestral population II. As in the case of traditional admixture analysis (16), we assume that the admixture has taken place in a single event at generation 0. If a case–parental control study was conducted at the $(t + 1)$th generation (parents from the tth generation), we have relative biases of

$$\mathrm{RB}_1^{(t+1)} = (1-2\theta) \cdot \delta^{(t)} \cdot [Q_{aB}^{(t)} R_1^{\mathrm{B}}(R_2^{\mathrm{B}}-1) + Q_{ab}^{(t)}(R_1^{\mathrm{B}}-1)]/F_1^{(t)} = (1-2\theta) \cdot (1-\theta)^t \cdot \delta^{(0)} \cdot [Q_{aB}^{(t)} R_1^{\mathrm{B}}(R_2^{\mathrm{B}}-1) + Q_{ab}^{(t)}(R_1^{\mathrm{B}}-1)]/F_1^{(t)}, \quad (7)$$

$$\mathrm{RB}_2^{(t+1)} = (1-2\theta) \cdot \delta^{(t)} \cdot [Q_{AB}^{(t)} R_1^{\mathrm{B}}(R_2^{\mathrm{B}}-1) + Q_{Ab}^{(t)}(R_1^{\mathrm{B}}-1)]/F_2^{(t)} = (1-2\theta) \cdot (1-\theta)^t \cdot \delta^{(0)} \cdot [Q_{AB}^{(t)} R_1^{\mathrm{B}}(R_2^{\mathrm{B}}-1) + Q_{Ab}^{(t)}(R_1^{\mathrm{B}}-1)]/F_2^{(t)}, \quad (8)$$

and, with a gene-dose effect for B,

$$\mathrm{RB}_1^{(t+1)} = \mathrm{RB}_2^{(t+1)} = (1-2\theta) \cdot (1-\theta)^t \cdot \delta^{(0)} \cdot (R^{\mathrm{B}}-1)/F^{(t)}. \quad (9)$$

The $\delta^{(0)}$ (linkage disequilibrium at generation 0) in these equations deserves special attention. It is composed of two terms, the first being the weighted average of the disequilibrium parameters of the two ancestral populations, and the second, the disequilibrium arising from population structure (16). That is, $\delta^{(0)} = [m\delta^{\mathrm{I}} + (1-m)\delta^{\mathrm{II}}] + [m(1-m)(P^{\mathrm{I}}{}_{A\bullet} - P^{\mathrm{II}}{}_{A\bullet})(P^{\mathrm{I}}{}_{\bullet B} - P^{\mathrm{II}}{}_{\bullet B})]$, where $P^{\mathrm{I}}{}_{A\bullet}$ and $P^{\mathrm{II}}{}_{A\bullet}$ denote the A allele frequencies in I and II, respectively, and $P^{\mathrm{I}}{}_{\bullet B}$ and $P^{\mathrm{II}}{}_{\bullet B}$ denote the corresponding values for B. Note that there is no confusion between the admixed population at $t = 0$ (genetic mixing allowed) and the previously defined stratified population (no genetic mixing).

In the admixed population, either the unlinked condition ($\theta = 0.5$) or the null-B condition ($R_1^{\mathrm{B}} = R_2^{\mathrm{B}} = 1$) will guarantee the unbiased estimation of GRRs of the A gene. However, a non null linked gene could by all means confound the GRR estimation, even if the gene is located very far away from the study gene. This is due to the disequilibrium arising from population structure ($\delta^{(t)} \neq 0$ even if $\delta^{\mathrm{I}} = \delta^{\mathrm{II}} = 0$, provided that the admixture is recent (t is not too large) and that there are frequency differences for both A and B genes in the two ancestral populations ($P^{\mathrm{I}}{}_{A\bullet} \neq P^{\mathrm{II}}{}_{A\bullet}$ and $P^{\mathrm{I}}{}_{\bullet B} \neq P^{\mathrm{II}}{}_{\bullet B}$)). Note that such a gene–gene confounding due to population structure does not appear, as shown in the previous section, in a stratified population without genetic mixing between its subpopulations.

Examples

Suppose a researcher conducted a case–parental control study for breast cancer in an unstructured population to estimate the GRRs for the *GSTT1* gene. The alleles of the *GSTT1* gene were grouped into "null" (high risk) and "non null" (low risk) groups, with frequency of the null alleles being about 0.4. However, the potential confounding effect from a nearby gene, the *COMT* gene, has been overlook in the study. (The *COMT* gene can be grouped into two alleles, "low activity" (high risk) and "high activity" (low risk), with rough equal frequency. The GRRs for *COMT* are $R_1^{\mathrm{COMT}} \approx R_2^{\mathrm{COMT}} \approx 2.5$). The genetic distance between these two genes is $\theta \approx 0.9\%$. Assume that the linkage disequilibrium between these two genes is $\delta = 0.12$ (Lewontin disequilibrium parameter (17) $D' = 0.6$), this study will then have relative biases of (using the above equations for unstructured population) $\mathrm{RB}_1 = \mathrm{RB}_2 = 50.6\%$. Such magnitude of bias deserves epidemiologists' full attention.

As another example, suppose a case–parental control study for breast cancer was conducted in an admixed population, this time to estimate the GRRs for the *SNCG* gene (at 10q23.2–10q23.3). The researcher however did not take into account the potential confounding effect from the *CYP17* gene, which is located at the same chromosome as (but far away from) the *SNCG* gene ($\theta \approx 15\%$). The GRRs for the *CYP17* gene are $R_1^{\mathrm{CYP17}} \approx R_2^{\mathrm{CYP17}} \approx 3$. Assume that

AEP Vol. 13, No. 4
April 2003: 261–266

the admixed population obtains equal proportion of its genes from two ancestral populations, I and II. The *SNCG* and *CYP17* genes are in linkage equilibrium in both ancestral populations (since the two genes are wide apart). The frequencies of the high-risk allele for *SNCG* gene in population I and II are, respectively, 0.8 and 0.2. And the corresponding values for *CYP17* gene are 0.7 and 0.1. If the study were conducted long after the admixture process (say, at the 30th generation), there will be relative biases of (using the above equations for admixed population) $RB_1^{(30)} = RB_2^{(30)} = 0.3\%$, which is negligible of course. However, if the study were conducted in a population that was recently admixed (say, at the 3rd generation), the relative biases will then amount to $RB_1^{(3)} = RB_2^{(3)} = 22.5\%$.

Table 1 shows the values of relative bias under various conditions, assuming a gene-dose effect for the confounder gene.

DISCUSSION

A comparison of our approach in this study to the association-mapping approach in the genetics literature (18–20) is in order, especially since both approaches use the same data structure of case–parents triads. For the mapping approach, interests are centered on using a marker (or markers) to "map" a putative susceptibility gene by linkage-disequilibrium tests. Naturally, the α level and the power of the test are the focal points. In our approach the genomic locations of the genes are assumed to be already known. What remains is then to quantify correctly the separate effects of the various genes on disease risk. And thus we encounter the confounder–gene problem. To make a sharper contrast between the two approaches, take the case of population admixture for example. Unlike the very narrow range ($<\sim 1$ cM) of disequilibrium likely seen in an unstructured population, an "admixture disequilibrium" can extend over a very long distance (or even the entire chromosome, if immediately after the admixture). This has since become the basis for the "admixture mapping" (21, 22) using the "transmission/disequilibrium test" (TDT) (19). In this case, admixture actually boosts the power of a TDT. However, our focus here is on estimation rather than on testing. And we found out that from the standpoint of gene–gene confounding, an admixture population is not a blessing but could be an origin for biases.

In conventional "risk factor" epidemiology, susceptibility genes were often treated just like any other exposure in the study, such as smoking, drinking, educational status, etc. However, our analysis has shown that as far as confounding effects are concerned, genes are very different from ordinary exposures. To determine whether another gene could confound the effects of the gene under concern, one must pay attention to the history of the study population (whether it is a recently admixed population, and how recent?). One must also consider the relative genomic positions of the two genes (whether the two genes are in the same chromosome, and how close?). By contrast, there is no such corresponding concept as "linkage" and/or "disequilibrium" between, say, smoking and drinking.

TABLE 1. Relative bias under various conditions, assuming a gene-dose effect for the confounder gene

Recombination fraction (%)	Lewontin disequilibrium parameter	Genotype relative risk for the confounder gene	Relative bias (%) for the gene under study			
			Unstructured population	Stratified* population	Admixed population	
					Generation = 3	Generation = 30
0.1	0.8	10	333.1	285.1	453.0	415.9
0.1	0.8	2	59.1	50.7	67.8	65.4
0.1	−0.8	10	−76.9	−29.7	58.4	56.4
0.1	−0.8	2	−37.1	−13.2	18.0	17.5
0.1	0.8	1	0.0	0.0	0.0	0.0
1.0	0.2	10	36.2	28.8	145.4	94.5
1.0	0.2	2	11.9	10.2	36.3	26.5
1.0	−0.2	10	−26.6	−8.7	89.1	61.4
1.0	−0.2	2	−10.6	−3.5	25.4	18.8
10.0	0	10	0.0	0.0	59.1	2.7
10.0	0	2	0.0	0.0	18.2	1.0
20.0	0	10	0.0	0.0	31.3	0.1
20.0	0	2	0.0	0.0	10.4	0.0
50.0	0	10	0.0	0.0	0.0	0.0

In the unstructured population, the frequency of the gene under study is 0.5, while that of the confounder gene is 0.4. The stratified (admixed) population are composed of (derived from) two subpopulations (source populations), I and II. The frequencies of the study gene are 0.8 (I) and 0.2 (II), while those of the confounder gene are 0.7 (I) and 0.1 (II). The Lewontin disequilibrium parameters are assumed equal in I and II.

* The two relative biases are in general not equal. Here the geometric average is presented.

From the derived relative-bias formula, we found that the degree of bias resulting from gene–gene confounding can be quite substantial under certain conditions (two genes are very close-linked and/or the study was performed in a recent admixed population). With epidemiology fast moving into a post-genomic era (23–25), one will be expecting more and more encounters of such a bias in real practices. Should that happen, we should consider jointly all the susceptibility genes in the same chromosome and avoid the taken-for-granted one-gene-at-a-time approach. The methodological details of the confounder-gene adjustment in case–parental control studies will be reported elsewhere.

This study was partly supported by a grant from the National Science Council, Republic of China.

REFERENCES

1. Ninkina NN, Alimova-Kost MV, Paterson JW, Delaney L, Cohen BB, Imreh S, et al. Organization, expression and polymorphism of the human persyn gene. Hum Mol Genet. 1998;7:1417–1424.
2. Dunning AM, Healey CS, Pharoah PD, Teare MD, Ponder BA, Easton DF. A systematic review of genetic polymorphisms and breast cancer risk. Cancer Epidem Biomar. 1999;8:843–854.
3. Schultz LB, Weber BL. Recent advances in breast cancer biology. Curr Opin Oncol. 1999;11:429–434.
4. Arver B, Du Q, Chen J, Luo L, Lindblom A. Hereditary breast cancer: a review. Semin Cancer Biol. 2000;10:271–288.
5. Moysich KB, Freudenheim JL, Baker JA, Ambrosone CB, Bowman ED, Schisterman EF, et al. Apolipoprotein E genetic polymorphism, serum lipoproteins, and breast cancer risk. Mol Carcinogen. 2000; 27:2–9.
6. Weber BL, Nathanson KL. Low penetrance genes associated with increased risk for breast cancer. Eur J Cancer. 2000;36:1193–1199.
7. Sun F, Flanders WD, Yang Q, Khoury MJ. A new method for estimating the risk ratio in studies using case-parental control design. Am J Epidemiol. 1998;148:902–909.
8. Lee WC, Chang CH. Estimating genotype relative risks in case-parental control studies: an optimal weighting approach. Am J Epidemiol. 2000;152:487–492.
9. Schaid DJ, Sommer SS. Genotype relative risks: methods for design and analysis of candidate-gene association studies. Am J Hum Genet. 1993;53:1114–1126.
10. Knapp M, Wassmer G, Baur MP. The relative efficiency of the Hardy-Weinberg equilibrium-likelihood and the conditional on parental genotype-likelihood methods for candidate-gene association studies. Am J Hum Genet. 1995;57:1476–1485.
11. Ewens WJ, Spielman RS. The transmission/disequilibrium test: history, subdivision and admixture. Am J Hum Genet. 1995;57:455–464.
12. Witte JS, Gauderman WJ, Thomas DC. Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. Am J Epidemiol. 1999; 149:693–705.
13. Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. J Natl Cancer Inst. 2000;92:1151–1158.
14. Greenland S. A unified approach to the analysis of case-distribution (case-only) studies. Stat Med. 1999;18:1–15.
15. Yang Q, Khoury MJ, Sun F, Flanders WD. Case-only design to measure gene-gene interaction. Epidemiology. 1999;10:167–170.
16. Chakraborty R, Smouse PE. Recombination of haplotypes leads to biased estimates of admixture proportions in human populations. Proc Natl Acad Sci USA. 1988;85:3071–3074.
17. Devlin B, Risch N. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics. 1995;29:311–322.
18. Ott J. Statistical properties of the haplotype relative risk. Genet Epidemiol. 1989;6:127–130.
19. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet. 1993;52:506–516.
20. Schaid DJ. General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol. 1996;13:423–49.
21. McKeigue PM. Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. Am J Hum Genet. 1997;60:188–196.
22. Kaplan NL, Martin ER, Morris RW, Weir BS. Marker selection for the transmission/disequilibrium test, in recently admixed populations. Am J Hum Genet. 1998;62:703–712.
23. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996;273:1516–1517.
24. Khoury MJ, Yang Q. The future of genetic studies of complex human diseases: an epidemiologic perspective. Epidemiology. 1998;9:350–354.
25. Khoury MJ, Little J. Human genome epidemiologic reviews: the beginning of something HuGE. Am J Epidemiol. 2000;151:2–3.

APPENDIX

Because the B gene was not observed, the data was classified according to the A gene (a total of ten categories of case-parents triads). Following the principle of Schaid (20), the number of triads in each of the ten categories can be derived (using the same notations as in text):

$$\#(AA \times AA \rightarrow AA) = N R_0 R_2^A R_1^A \cdot (P_A^2 . P_{AB}^2 R_2^B R_1^B + 2P_A^2 . P_{AB} P_{Ab} R_1^B + P_A^2 . P_{Ab}^2)' \quad (10)$$

$$\#(AA \times Aa \rightarrow AA) = 2N R_0 R_2^A R_1^A \cdot \{P_A . P_{AB}(P_a . P_{AB} - \theta\delta) R_2^B R_1^B + [P_A . P_{AB} (P_a . P_{Ab} + \theta\delta) + P_A . P_{Ab}(P_a . P_{AB} - \theta\delta)] R_1^B + P_A . P_{Ab}(P_a . P_{Ab} + \theta\delta)\}, \quad (11)$$

$$\#(AA \times Aa \rightarrow Aa) = 2N R_0 R_1^A \cdot \{P_A . P_{AB}(P_A . P_{aB} + \theta\delta) R_2^B R_1^B + [P_A . P_{AB} (P_A . P_{ab} - \theta\delta) + P_A . P_{Ab}(P_A . P_{aB} + \theta\delta)] R_1^B + P_A . P_{Ab}(P_A . P_{ab} - \theta\delta)\}, \quad (12)$$

$$\#(AA \times aa \rightarrow Aa) = N R_0 R_1^A \cdot [2P_A . P_a . P_{AB} P_{aB} R_2^B R_1^B + 2P_A . P_a . (P_{AB} P_{ab} + P_{Ab} P_{aB}) R_1^B + 2P_A . P_a . P_{Ab} P_{ab}], \quad (13)$$

AEP Vol. 13, No. 4
April 2003: 261–266

$$\#(Aa \times Aa \to AA) = N R_0 R_2^A R_1^A \cdot [(P_{a.}P_{AB} - \theta\delta)^2 R_2^B R_1^B + 2(P_{a.}P_{AB} - \theta\delta)(P_{a.}P_{Ab} + \theta\delta)R_1^B + (P_{a.}P_{Ab} + \theta\delta)^2], \quad (14)$$

$$\#(Aa \times Aa \to Aa) = 2NR_0 R_1{}^A \cdot \left\{(P_{a.}P_{AB} - \theta\delta)(P_{A.}P_{aB} + \theta\delta)R_2^B\ R_1^B + [(P_{a.}P_{AB} - \theta\delta)(P_{A.}(P_{ab} - \theta\delta) + (P_{a.}P_{Ab} + \theta\delta)(P_{A.}P_{aB} + \theta\delta)]R_1^B + (P_{a.}P_{Ab} + \theta\delta)(P_{A.}P_{ab} - \theta\delta)\right\}, \quad (15)$$

$$\#(Aa \times Aa \to aa) = NR_0 \cdot [(P_{A.}P_{aB} + \theta\delta)^2 R_2^B\ R_1^B + 2(P_{A.}P_{aB} + \theta\delta)(P_{A.}P_{ab} - \theta\delta)R_1^B + (P_{A.}P_{ab} - \theta\delta)^2], \quad (16)$$

$$\#(Aa \times aa \to Aa) = 2NR_0 R_1^A \cdot \left\{P_{a.}P_{aB}(P_{a.}P_{AB} - \theta\delta)R_2^B\ R_1^B + [P_{a.}P_{aB}(P_{a.}P_{Ab} + \theta\delta) + P_{a.}P_{ab}(P_{a.}P_{AB} - \theta\delta)]R_1^B + P_{a.}P_{ab}(P_{a.}P_{Ab} + \theta\delta)\right\}, \quad (17)$$

$$\#(Aa \times aa \to aa) = 2NR_0 \cdot \left\{P_{a.}P_{aB}(P_{A.}P_{aB} + \theta\delta)R_2^B\ R_1^B + [P_{a.}P_{aB}(P_{A.}P_{ab} - \theta\delta) + P_{a.}P_{ab}(P_{A.}P_{aB} + \theta\delta)]R_1^B + P_{a.}P_{ab}(P_{A.}P_{ab} - \theta\delta)\right\}, \quad (18)$$

$$\#(aa \times aa \to aa) = NR_0 \cdot (P_{a.}^2\ P_{aB}^2\ R_2^B\ R_1^B + 2P_{a.}^2\ P_{aB}P_{ab}R_1^B + P_{a.}^2\ P_{ab}^2). \quad (19)$$

The GRR estimates for the A gene are (7):

$$\hat{R_1^A} = \frac{\#(Aa \times aa \to Aa) + \#(Aa \times Aa \to Aa)}{\#(Aa \times aa \to aa) + \text{twice of } \#(Aa \times Aa \to aa)} \quad (20)$$

and

$$\hat{R_2^A} = \frac{\#(AA \times Aa \to AA) + \text{twice of } \#(Aa \times Aa \to AA)}{\#(AA \times Aa \to Aa) + \#(Aa \times Aa \to Aa)} \quad (21)$$

(If desired, one can use the Lee and Chang's method (8) to achieve greater precision in GRR estimation. Here we resort to the simpler method of Sun et al. (7), since the focus of this paper is on the anatomy of bias but not on precision.)

The RBs for the A gene are then (after some algebra):

$$\mathrm{RB}_1 = \left(\hat{R_1^A} - R_1^A\right)/R_1^A = (1 - 2\theta) \cdot \delta \cdot [Q_{aB}R_1^B\ (R_2^B - 1) + Q_{ab}(R_1^B - 1)]/F_1 \quad (22)$$

and

$$\mathrm{RB}_2 = \left(\hat{R_2^A} - R_2^A\right)/R_2^A = (1 - 2\theta) \cdot \delta \cdot [Q_{AB}R_1^B\ (R_2^B - 1) + Q_{Ab}(R_1^B - 1)]/F_2, \quad (23)$$

where

$$F_1 = Q_{aB}(P_{A.}P_{aB} + \theta\delta)R_2^B\ R_1^B + [Q_{ab}(P_{A.}P_{aB} + \theta\delta) + Q_{aB}(P_{A.}P_{ab} - \theta\delta)]R_1^B + Q_{ab}(P_{A.}P_{ab} - \theta\delta) \quad (24)$$

and

$$F_2 = Q_{AB}(P_{A.}P_{aB} + \theta\delta)R_2^B\ R_1^B + [Q_{Ab}(P_{A.}P_{aB} + \theta\delta) + Q_{AB}(P_{A.}P_{ab} - \theta\delta)]R_1^B + Q_{Ab}(P_{A.}P_{ab} - \theta\delta). \quad (25)$$

When $R_1{}^B = R_2{}^B = R^B$, the above equations for RBs reduce to

$$\mathrm{RB}_1 = \mathrm{RB}_2 = (1 - 2\theta) \cdot \delta \cdot (R^B - 1)/F, \quad (26)$$

where

$$F = (P_{A.}P_{aB} + \theta\delta)R^B + (P_{A.}P_{ab} - \theta\delta). \quad (27)$$